Protein Engineering vol.4 no 3 pp.359–370, 1991

Corrigendum

When the following paper was originally published on pages 137–147 of this volume, part of Figure 2 was omitted. The paper is reprinted here in full.

Improving the prediction of secondary structure of 'TIM-barrel' enzymes

Thomas Niermann and Kasper Kirschner[1]

Abteilung Biophysikalische Chemie, Biozentrum, University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland

[1]To whom correspondence should be addressed

The information contained in aligned sets of homologous protein sequences should improve the score of secondary structure prediction. Seven different enzymes having the $(\beta/\alpha)_8$ or TIM-barrel fold were used to optimize the prediction with regard to this class of enzymes. The α-helix, β-strand and loop propensities of the Garnier–Osguthorpe–Robson method were averaged at aligned residue positions, leading to a significant improvement over the average score obtained from single sequences. The increased accuracy correlates with the average sequence variability of the aligned set. Further improvements were obtained by using the following averaged properties as weights for the averaged state propensities: amphipathic moment and α-helix; hydropathy and β-strand; chain flexibility and loop. The clustering of conserved residues at the C-terminal ends of the β-strands was used as an additional positive weight for β-strand propensity and increased the prediction of otherwise unpredicted β-strands decisively. The automatic weighted prediction method identifies >95% of the secondary structure elements of the set of seven TIM-barrel enzymes.

Key words: $(\beta/\alpha)_8$ barrel/homologous proteins/secondary structure prediction/TIM barrel

Introduction

It is desirable to improve the accuracy of empirically predicting the secondary structure of protein sequences. Progress in these efforts could ultimately lead to the correct prediction of the chain fold (Taylor, 1987, 1988; Argos and McCaldon, 1988; Schulz, 1988; Argos, 1990).

The empirical prediction algorithms rarely achieve >60% correctly predicted residue positions. Joint predictions involving the combination of several different empirical methods slightly alleviate this unsatisfactory situation (Biou *et al.*, 1988). Further improvements should result from exploiting the redundant folding information contained in homologous protein sequences (Wooton, 1974; Chothia and Lesk, 1987), which are becoming increasingly available due to the development of cDNA cloning and sequencing techniques. Since the secondary structure of small, single-domain monomeric proteins is predicted with greater accuracy than that of large, multi-domain oligomers (Argos *et al.*, 1976), it appears advisable to concentrate on the former class of proteins.

We have recently used 10 different amino acid sequences of the α subunit of tryptophan synthase to predict its secondary structure (Crawford *et al.*, 1987). Two different empirical prediction methods (Chou and Fasman, 1978; Garnier *et al.*, 1978) were used on the aligned set of sequences, and the consensus predictions were combined with patterns of averaged hydrophobicity and chain flexibility to achieve a joint prediction. The sequence of secondary structure elements was consistent with the chain fold of an 8-fold $\beta\alpha$ [$(\beta/\alpha)_8$ or TIM]-barrel. Data on the limited proteolysis, chemical modification and mutagenesis of the protein from *Escherichia coli* further supported the suggested protein fold, which was subsequently established by protein crystallography (Hyde *et al.*, 1988). At least 18 different enzymes have the $(\beta/\alpha)_8$-barrel fold (Farber and Petsko, 1990). Figure 1 shows indoleglycerol phosphate synthase (IGPS; Priestle *et al.*, 1987) as an example. This structure consists of a single structural domain in which the alternating β-strands and α-helices interact mainly with their neighbours and where quaternary interactions do not seem to be important for stability. In the work reported here we have chosen those seven $(\beta/\alpha)_8$-barrel enzymes that are presently associated with seven or more known amino acid sequences from different organisms to develop an automatic procedure for improved prediction of their secondary structure.

We show that, at a first level, averaging of the empirical predictions of secondary structure, according to Garnier *et al.* (1978) and Gibrat *et al.* (1987), at each aligned residue position improves the score in proportion to the variability of the sequence. Moreover, the patterns of averaged amphipathic helical moment, hydrophobicity and chain flexibility can be used as weights for improving the score further. At a second level the correlation between β-strands and segments of low sequence variability was used to improve specifically the prediction of β-strands. Finally,

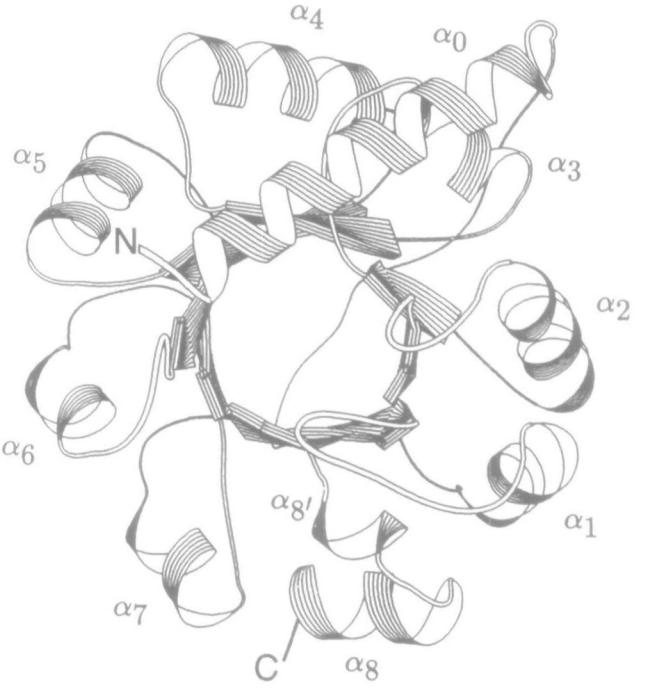


Fig. 1. Schematic ribbon diagram of a $(\beta/\alpha)_8$-barrel enzyme. IGPS synthase from *E.coli*. The C-terminal ends of the internal β-strands (arrows) in the barrel are pointing towards the reader. The external α-helices are numbered sequentially from the N-terminus.

it is shown that the 8-fold repeat of a specific template of supersecondary structure matches well to the observed sequences of secondary structure elements of the seven $(\beta/\alpha)_8$-barrel enzymes in the learning set.

Materials and methods

Alignment of amino acid sequences

The sequences of the proteins used in the work reported here were obtained from the MIPS data bank (Mewes, 1990). The sequences were first aligned in pairs, using the computer program FASTA (Pearson and Lipman, 1988; GCG Sequence Analysis Software Package, 1989). The computer programs PROFILE and PROFILEGAP (Gribskov *et al.*, 1988; GCG Sequence Analysis Software Package, 1989) were used for multisequence alignment. Minor rearrangements in the sequence alignments were greatly aided by the frequent occurrence of conserved residues at the C-terminus of β-strands or the following loops, or both. The final sequence alignment was carried out by eye, aligning the gaps and minimizing their number as far as possible. Gaps were located exclusively to known surface loops. The periodic clustering of both identical residues and gaps throughout the sequences provided the necessary internal register for confident alignment.

The numbering of the set begins with the first amino acid of that sequence which has the longest N-terminal extension (not shown here). It is continuous throughout the longest segments that span each of the clustered gaps. The variability at each residue position was defined according to Wu and Kabat (1970) and used for constructing the variability weighting profile as described in the text.

Averaging of secondary structure propensities

The secondary structure 'GOR' prediction algorithm (Garnier *et al.*, 1978; Gibrat *et al.*, 1987) was first used to calculate the propensity of each residue for one of the three states H (α-helix), S (β-strand) and C (non-H, non-S or coil, loop) for each original continuous amino acid sequence. Then the propensities corresponding to each residue were located to that residue's position in the matrix of aligned sequences. The average state propensities at each position were obtained by summing up and dividing the sum by the actual number of occupied positions in that column (Garnier *et al.*, 1978). The predictions were finally converted to averaged 'state profiles' by smoothing with a three-residue span. The positions corresponding to the inserts (which define the gaps in the alignments) are given the same weight as the fully occupied positions on both sides of the gap, but for smoothing purposes only. We preferred this procedure to the use of Gly as a dummy residue in gap positions (Zvelebil *et al.*, 1987). The state with the highest average propensity defines the predicted state at that position. This procedure is unbiased and unequivocal. By contrast, consensus procedures in which the secondary structure is determined by the predominant state can lead to ambiguous results (Crawford *et al.*, 1987).

Accuracy of the prediction and optimization of decision constants (DCs)

There are different ways to estimate the accuracy of a secondary structure prediction on a per-residue basis. It can be quantified either by suitable quotients (Kabsch and Sander, 1983) or by correlation coefficients (Matthews, 1975). The distribution of all predicted states over all known states constitutes a 3 × 3 scoring matrix (Schulz, 1988; see also Table I). Taylor and Thornton (1984) have shown that only three states (α-helix, H; β-strand, S; loop, C) need to be considered for α/β proteins (Levitt and Chothia, 1976).

The most commonly used measure of accuracy is the quotient Q_3 (the 'success rate'), i.e. the sum of the diagonal elements of the scoring matrix divided by N, the total number of residues considered.

$$Q_3 = \frac{H + S + C}{N}$$

N is smaller than given by the total number of residue positions in the aligned set of sequences. This difference results from ignoring gap positions and overhanging N- and C-termini, justified as follows. First, the available sequence data in gap regions is necessarily limited and second, there is no X-ray information on the secondary structure of segments which are inserted relative to the reference sequence.

Since the three different conformational states are not populated equally, there is an inherent tendency to bias Q_3 towards improved prediction of the more abundant H and C states at the expense of the S state. In this work we balanced the scores of the three conformational states H, S and C, expressed as Q_I values. For example, $Q_{I,H} = H/\Sigma_H$ where H is the number of correctly predicted H-states divided by Σ_H, the total of known H-states (i.e. the sum of the horizontal cells in the matrix of Table I). We observed that, upon varying the decision constants (Garnier *et al.*, 1978) for α-helix, DC(H) and β-strand, DC(S) in steps of 25 from −50 to 50 centinats [leaving DC(C) constant at 0], Q_3 values for each of the seven test proteins varied only in a limited fashion, but the score for $Q_{I,S}$ varied strongly.

If the accuracy of prediction were determined on a per-element basis, a relatively large number of β-strands would be unpredicted due to a relatively small fraction of unpredicted single β-strand residues. For the combination of DC(H) = 50 and DC(S) = −25 the Q_I values were approximately equal for all three states in six out of seven test proteins. For alpha amylase (Amy) the optimal combination was DC(H) = 50 and DC(S) = 25. By contrast the DCs that were optimized by Taylor and Thornton (1984), and later confirmed by Gibrat *et al.* (1987), were also found to be optimal for the prediction of single sequences used here. The data on the average score of single sequences given in Tables I and II were obtained with the values of Gibrat *et al.* (1987), DC(H) = 25, DC(S) = 30 and DC(C) = 0.

We used the structure abstract procedure of Taylor (1984), for scoring predicted strings of secondary structure elements with respect to the known secondary structure (see Table II). The following additional rules provide for an unambiguous quantification of the prediction on a per-element basis. Predicted elements were classified into true positive (tp), false positive (fp) and false negative (fn). The sum (tp + fn) corresponds to the total of known H(S) elements. True negatives (tn) were not considered because they coincide with the category 'tp' of the other state prediction. Predictions within clustered gap positions were converted to coil predictions. The minimal length of a predicted helix was either four continuous H predictions or an interrupted run of four H out of five positions. The minimal length of a predicted β-strand was either three continuous S predictions or an interrupted run of three S out of four positions. Runs of only three α-helical or two β-strand positions were eliminated by reassigning them evenly to the state of the positions at the two sides. The classification of the turn or coil prediction was omitted because 'fp' coils appear as 'fn' in the H or S prediction and 'fn' coils appear as 'fp' in the H or S prediction. Predicted H(S) elements were classified as 'tp' if they overlapped with known H(S) secondary structure elements by at least two residues. Missing overlaps of known H(S) elements with predicted H(S) elements were

classified as 'fn'. Predicted H(S) elements were classified as 'fp' if they overlapped with known S(H) secondary structure elements by at least two residues. This rule was applied with the following exception: if a predicted H(S) element qualified as 'tp' but also overlapped an adjacent known S(H) element, this second overlap was only classified 'fp' if it was not simultaneously overlapped by a 'tp' S(H) prediction. The intention was to ignore incorrect overlaps where the sequence of secondary structure elements was predicted correctly. This exception had to be considered only once (see the predictions for β_8, α'_8 of aldolase in Figure 4), but it poses a general problem for the quantitative scoring of predicted secondary structure elements. Table II shows the per-element score of each set of aligned sequences, which was evaluated from the data presented in Figure 4.

Profiles of averaged residue properties

The profiles of chain flexibility of individual sequences were calculated according to Karplus and Schulz (1985). Hydropathy profiles are calculated using the scale of Kyte and Doolittle (1982) over a five-residue span setting, and were finally smoothed with a three-residue span. The scale of Kyte and Doolittle (1982) (Cornette *et al.*, 1987) was also used for the calculation of the helical amphipathic moment at 100 degrees (Eisenberg *et al.*, 1984) with a span setting of 11 residues.

The chosen span settings are reasonable with respect to the average lengths of observed secondary structure elements; α-helix: $n \simeq 12$; β-strand: $n \simeq 5$ (Taylor and Thornton, 1984). The averaging of these 'property profiles' for an aligned set of sequences followed essentially the procedures described above for the state profiles.

State prediction weighted by property

The three property profiles were first normalized to a maximum value of 600 centinats (Garnier *et al.*, 1978) to allow quantitative weighting with the 'state profiles'. Then a threshold of 150 centinats was defined by iterative optimization. If a given property profile exceeded this threshold it was added to the corresponding state profile and the sum divided by two to give the 'weighted averaged state profile' in that region of sequence. The geometric mean gave practically the same improvements.

β-Strand prediction reinforced by sequence variability

Wu and Kabat (1970) defined sequence variability as follows. The number of different amino acids at a given position of aligned sequences is divided by the fraction of the predominant residue. This variability parameter (line var. in Figure 2A) was used to enhance the β-strand prediction in regions of low variability. The following procedure is the result of empirical optimization. The value of the normalized average β-strand propensity was increased by 100 centinats only if the two following conditions were fulfilled simultaneously: (i) the average variability over a span of five residues was $<30\%$ of the average variability parameter over the entire set of aligned sequences; and (ii) both the average hydropathy parameters in this region were > -150 centinats and the average flexibility parameters were <150 centinats. The second condition prevented overprediction of β-strands in the subsequent loop regions.

Results

$(\beta/\alpha)_8$-Barrel enzymes

The seven $(\beta/\alpha)_8$-barrel enzymes of the learning set were phosphoribosyl anthranilate isomerase (PRAI); indoleglycerol phosphate synthase (IGPS); the α-subunit of tryptophan synthase (TSA); triose phosphate isomerase (TIM); aldolase (Aldo); alpha amylase (Amy) and enolase (Eno). Figure 1 is a schematic ribbon diagram (Priestle, 1988) of one member of the learning set, IGPS from *Escherichia coli* (Priestle *et al.*, 1987). Except for the additional α-helix α_0 at the N-terminus, and the short α-helix $\alpha_{8'}$ in the loop between β_8 and α_8 IGPS is a 'limit $(\beta/\alpha)_8$-barrel' protein. The eight parallel β-strands that are buried in the centre are connected on the outside by eight α-helices that are antiparallel to the eight β-strands. Enolase is an unorthodox $\beta_1\beta_2\alpha_1\alpha_2$ $(\beta/\alpha)_6$-barrel enzyme (Lebioda *et al.*, 1989). The alignments and the predictions were performed with the full sequences of each protein. Alpha amylase is the exception, where both the large B-domain that is inserted between β_3 and α_3 and the C-terminal domain were eliminated. The B-domain was treated as a gap for the quantitative evaluation. Although the full sequence of enolase was aligned and analysed, only the β-barrel domain was considered for the quantitative evaluation.

Sequence alignment of indoleglycerol phosphate synthase

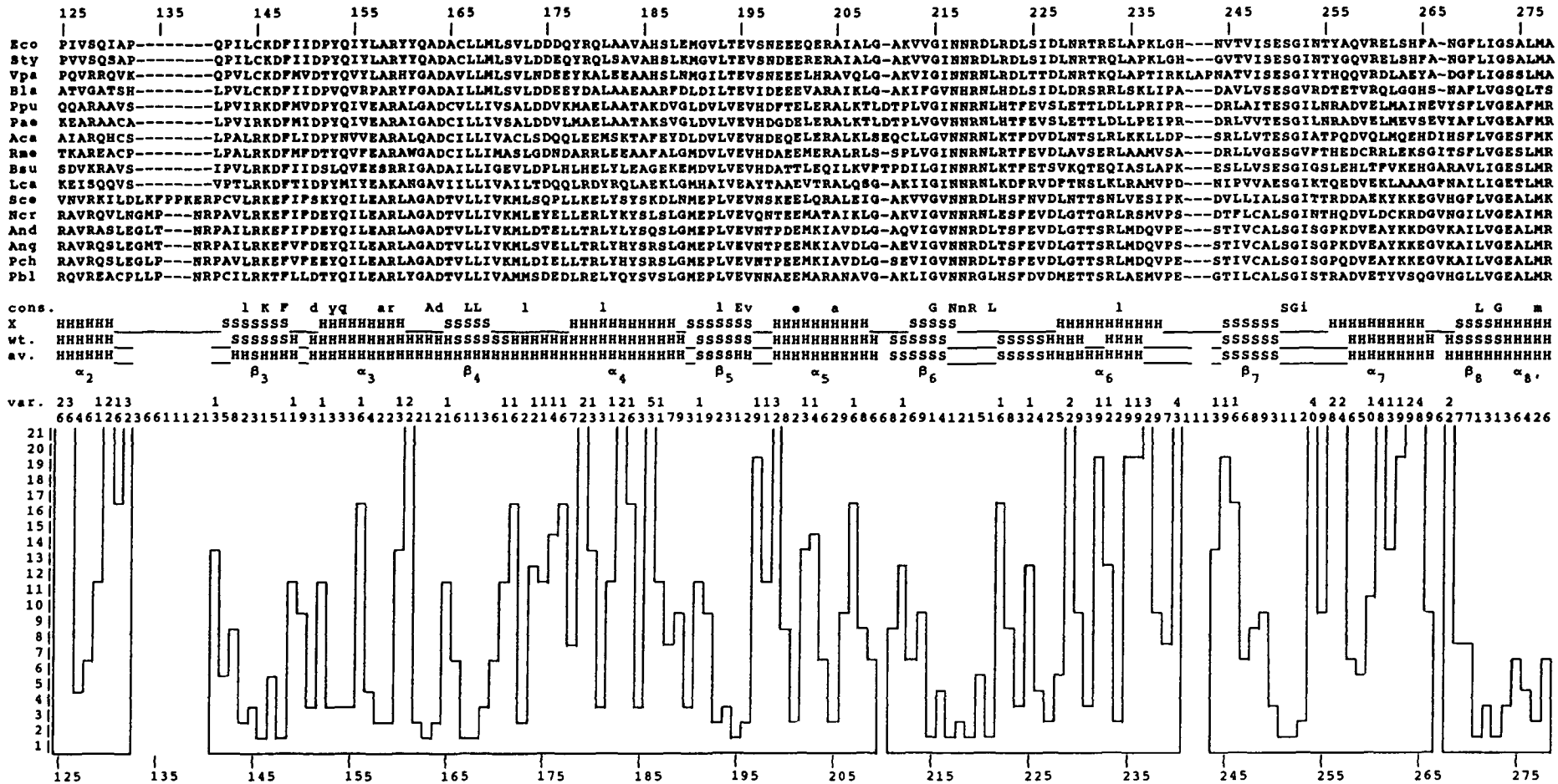
Figure 2(A) presents a region, 150 amino acids long, of a set of aligned sequences of IGPS from 16 different organisms. This region was selected because it is one of the least-well-predicted ones in the set of seven test proteins. It comprises $\sim 80\%$ of the secondary structure elements of IGPS (see Figure 1). The general approach of averaging and weighting of state propensities will be illustrated qualitatively with this representative region. The quantitative prediction scores involving the full set of seven $(\beta/\alpha)_8$-barrel enzymes will be presented further on.

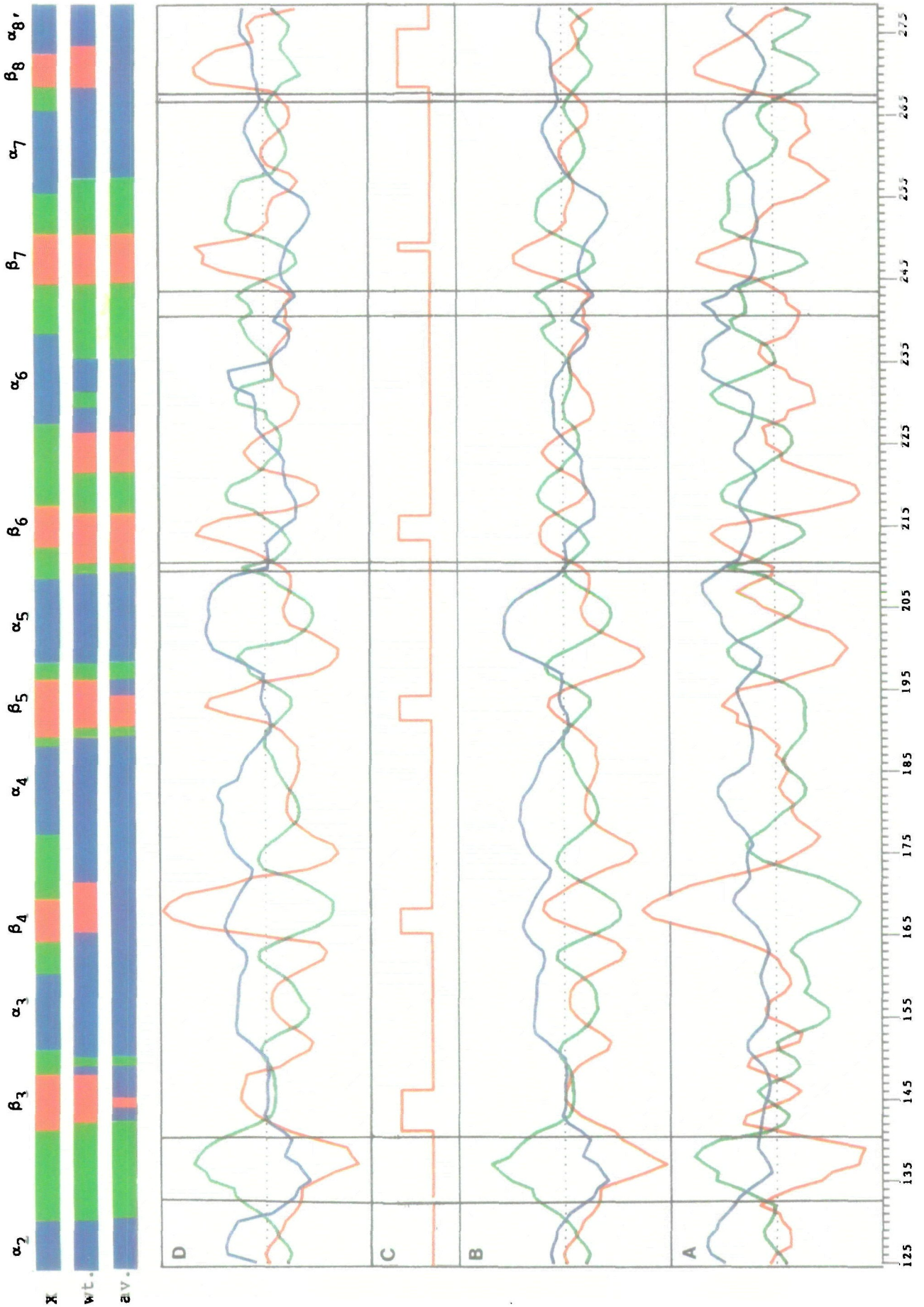
The first row below the stack of aligned sequences indicates conserved residues. Upper-case letters correspond to 16 identical residues, whereas lower-case letters allow for two nonidentical residues, anticipating possible sequencing errors. Several of these residues are candidates for either substrate binding or catalytic function or both (e.g. Glu195, Asn217, Arg219 and Ser251). Residue positions in the vicinity of conserved ones frequently contain residues of similar polarity and size. The variability of the amino acid sequence is expressed by the histogram of the variability parameter defined by Wu and Kabat (1970), on a truncated scale (bottom of Figure 2A).

The next row ('X') further down corresponds to the known secondary structure of IGPS from *E.coli* (Priestle *et al.*, 1987; M.Wilmanns and J.N.Jansonius, personal communication), comprising all secondary structure elements of IGPS between β_2 and α_8 (see Figure 1). Blank spaces correspond to clustered gaps in the set of aligned sequences above.

Averaged secondary structure prediction

Garnier *et al.* (1987) proposed that homologous sequences could be used to improve the accuracy of secondary structure prediction by averaging. Panel B of Figure 2(B) depicts the three averaged secondary 'state profiles' of the aligned IGPS sequences of Figure 2(A). The coloured bar codes at the top of Figure 2(B) represent the sequences of secondary structure elements defined by the corresponding profiles below. Here, blue is α-helix, red is β-strand and green is loop. Thus the bar code labelled 'av' corresponds to panel B, whereas the upper bar code labelled 'X' corresponds to the known secondary structure of IGPS. These known and predicted secondary structure elements are repeated in Figure 2(A), in the row labelled 'X', as runs of H (α-helix), S (β-strand), the underscore character _ (loop) and blank spaces (gap) below the set of aligned amino acid sequences. Comparison between the bar codes labelled 'X' and 'av' shows that β_3 is underpredicted, and β_4 is incorrectly predicted as a portion of one long α-helix (position 150–190). β_8 is also mispredicted as α-helix, and a loop segment around position 223 is mispredicted as β-strand.


125 135 145 155 165 175 185 195 205 215 225 235 245 255 265 275
Eco PIVSQIAP--------QPILCKDFIIDPYQIYLARYYQADACLLMLSVLDDDQYRQLAAVAHSLEMGVLTEVSNEEEQERAIALG-AKVVGINNRDLRDLSIDLNRTRELAPKLGH---NVTVISESGINTYAQVRELSHFA-NGFLIGSALMA
Sty PVVSQSAP--------QPILCKDFIIDPYQIYLARYYQADACLLMLSVLDDEQYRQLSAVAHSLKMGVLTEVSNDEERERAIALG-AKVVGINNRDLRDLSIDLNRTRQLAPKLGH---GVTVISESGINTYGQVRELSHFA-NGFLIGSALMA
Vpa PQVRRQVK--------QPVLCKDFMVDTYQVYLARHYGADAVLLMLSVLNDEEYKALEEAAHSLNMGILTEVSNEEELHRAVQLG-AKVIGINNRNLRDLTTDLNRTKQLAPTIRKLAPNATVISESGIYTHQQVRDLAEYA-DGFLIGSSLMA
Bla ATVGATSH--------LPVLCKDFIIDPVQVRPARYFGADAILLMLSVLDDEEYDALAAEAARFDLDILTEVIDEEEVARAIKLG-AKIFGVNHRNLHDLSIDLDRSRRLSKLIPA---DAVLVSESGVRDTETVRQLGGHS-NAFLVGSQLTS
Ppu QQARAAVS--------LPVIRKDFMVDPYQIVEARALGADCVLLIVSALDDVKMAELAATAKDVGLDVLVEVHDFTELERALKTLDTPLVGINNRNLHTFEVSLETTLDLLPRIPR---DRLAITESGILNRADVELMAINEVYSFLVGEAFMR
Pae KEARAACA--------LPVIRKDFMIDPYQIVEARAIGADCILLIVSALDDVLMAELAATAKSVGLDVLVEVHDGDELERALKTLDTPLVGVNNRNLHTFEVSLETTLDLLPEIPR---DRLVVTESGILNRADVELMEVSEVYAFLVGEAFMR
Aca AIARQHCS--------LPALRKDFLIDPYNVVEARALQADCILLIVACLSDQQLEEMSKTAFEYDLDVLVEVHDEQELERALKLSEQCLLGVNNRNLKTFDVDLNTSLRLKKLLDP---SRLLVTESGIATPQDVQLMQEHDIHSFLVGESFMK
Rme TKAREACP--------LPALRKDFMFDTYQVFEARAWGADCILLIMASLGDNDARRLEEAAFALGMDVLVEVHDAEEMERALRLS-SPLVGINNRNLRTFEVDLAVSERLAAMVSA---DRLLVGESGVFTHEDCRRLEKSGITSFLVGESLMR
Bsu SDVKRAVS--------IPVLRKDFIIDSLQVEESRRIGADAILLIGEVLDPLHLHELYLEAGEKEMDVLVEVHDATTLEQILKVFTPDILGINNRNLKTFETSVKQTEQIASLAPK---ESLLVSESGIGSLEHLTFVKEHGARAVLIGESLMR
Lca KEISQQVS--------VPTLRKDFTIDPYMIYEAKANGAVIILLIVAILTDQQLRDYRQLAEKLGMHAIVEAYTAAEVTRALQSG-AKIIGINNRNLKDFRVDFTNSLKLRAMVPD---NIPVVAESGIKTQEDVEKLAAAGFNAILIGETLMR
Sce VNVRKILDLKFPPKERPCVLRKEFIFSKYQILEARLAGADTVLLIVKMLSQPLLKELYSYSKDLNMEPLVEVNSKEELQRALEIG-AKVVGVNNRDLHSFNVDLNTTSNLVESIPK---DVLLIALSGITTRDDAEKYKKEGVHGFLVGEALMK
Ncr RAVRQVLNGMP---NRPAVLRKEFIFDEYQILEARLAGADTVLLIVKMLEYELLERLYKYSLSLGMEPLVEVQNTEEMATAIKLG-AKVIGVNNRNLESFEVDLGTTGRLRSMVPS---DTFLCALSGINTHQDVLDCKRDGVNGILVGEAIMR
And RAVRASLEGLT---NRPAILRKEFIFDEYQILEARLAGADTVLLIVKMLDTELLTRLYLYSQSLGMEPLVEVNTPDEMKIAVDLG-AQVIGVNNRDLTSFEVDLGTTSRLMDQVPE---STIVCALSGISGPKDVEAYKKDGVKAILVGEALMR
Ang RAVRQSLEGMT---NRPAILRKEFVFDEYQILEARLAGADTVLLIVKMLSVELLTRLYHYSRSLGMEPLVEVNTPEEMKIAVDLG-AEVIGVNNRDLTSFEVDLGTTSRLMDQVPS---STIVCALSGISGPKDVEAYKKEGVKAILVGEALMR
Pch RAVRQSLEGLP---NRPAVLRKEFVFEEYQILEARLAGADTVLLIVKMLDIELLTRLYHYSRSLGMEPLVEVNTPEEMKIAVDLG-SEVIGVNNRDLTSFEVDLGTTSRLMDQVPE---STIVCALSGISGPQDVEAYKKEGVKAILVGEALMR
Pbl RQVREACPLLP---NRPCILRKTFLLDTYQILEARLYGADTVLLIVAMMSDEDLRELYQYSVSLGMEPLVEVNNAEEMARANAVG-AKLIGVNNRGLHSFDVDMETTSRLAEMVPE---GTILCALSGISTRADVETYVSQGVHGLLVGEALMR
cons.
X
wt.
av.
α2 β3 α3 β4 α4 β5 α5 β6 α6 β7 α7 β8 α8
var.
21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
125 135 145 155 165 175 185 195 205 215 225 235 245 255 265 275


α2
β3
α3
β4
α4
β5
α5
β6
α6
β7
α7
β8
α8'
x
wt.
av.
D
C
B
A
125
135
145
155
165
175
185
195
205
215
225
235
245
255
265
275

Averaged prediction improves with increasing sequence variability

The degeneracy of the folding code (Jaenicke, 1987) (Figure 2A) that permits many different sequences to fold to the same secondary structure element in a given tertiary structure suggests that the relative improvement of prediction accuracy should depend on the variability of the sequence. To test this idea we compared the percentage increase of prediction accuracy to the average variability of each protein in the learning set. The following α/β proteins with chain folds that differ from the $(\beta/\alpha)_8$-barrel motif were included to check the generality of this proposal: aspartate aminotransferase (AAT) (Eichele *et al.*, 1979; J.N.Jansonius, personal communication) and dihydrofolate reductase (DHFR) (Matthews *et al.*, 1977). As seen in Figure 3, there is a clear correlation between the percentage increase of prediction accuracy obtained by averaging (Δ%) and the overall sequence variability. The accuracy cannot increase indefinitely. It appears that the relationship shown in Figure 3 approaches a plateau at a variability index above 14. A further increase of variability would render the sequence alignment progressively equivocal (Strasser *et al.*, 1989). Predictions were performed with the optimal decision constants DC(H) = 50, DC(S) = −25 and DC(C) = 0, as described in Materials and methods. For Amy, Q_3 increases by 8%, but in this case these constants lead to overprediction of β-strands, for unknown reasons. The empty symbol (□) for Amy in Figure 3 corresponds to an increase of Q_3 by 13%, which was obtained with DC(S) = 25, leaving DC(H) and DC(C) unchanged.

Averaged property profiles

Profiles of physico-chemical residue properties correlate with structural features and are useful to identify secondary structure elements. Patterns of residue properties are furthermore well-tried ingredients of pattern-matching methods (Lim, 1974; Cohen *et al.*, 1983). The previous study of TSA α-subunit (Crawford *et al.*, 1987) had shown that there is generally good agreement between buried β-strands and local maxima of averaged hydropathy profiles (Kyte and Doolittle, 1982) as well as between surface loops and local maxima of averaged chain flexibility profiles (Karplus and Schulz, 1985). Since then we have found that the local maxima of the averaged amphipathic helical moment (Eisenberg *et al.*, 1984) correlated well with the known α-helices in all seven test proteins. Panel A of Figure 2(B) presents the profiles of the averaged amphipathy, hydropathy and flexibility values corresponding to the representative region of IGPS in Figure 2(A). They were calculated and smoothed as described in Materials and methods.

It is seen that all known β-strands are associated with pronounced local maxima in the hydropathy profile, whereas the correlation with the incorrectly predicted β-strand around residue position 223 is only subliminal. The maxima in the flexibility profiles generally correlate well with known loop segments. Moreover the profiles of hydropathy and flexibility generally vary in a reciprocal fashion. Quantitative comparisons (omitted here) showed that maxima of β-strand profiles correlated better with maxima of hydropathy than with minima of chain flexibility. Similarly, maxima of loop profiles correlated better with maxima of chain flexibility than with minima of hydropathy. Finally, the amphipathy profiles, which have only positive values, correlate well with the known α-helical elements. Similar correlations were observed for the other test proteins (data not shown).

Weighted state prediction

Maxima of 'cognate' property profiles correlate well with known secondary structure elements. As already pioneered by Taylor and Thornton (1984) these qualitative correlations were used to develop an automatic weighting procedure, as follows. First, the three averaged property profiles were added to the corresponding averaged state profiles and the sum divided by two. The weighted averaged state profiles of α-helix, β-strand and loop are presented in panel D and in the corresponding colour bar code labelled 'wt' in Figure 2(B), and the symbol code labelled 'wt' in Figure 2(A).

Zvelebil *et al.* (1987) have shown recently that sequence variability profiles can be used to improve secondary structure prediction. The periodic decrease of variability that is observed at the C-terminus of β-strands and the subsequent loops is caused by the clustering of conserved residues required for function of

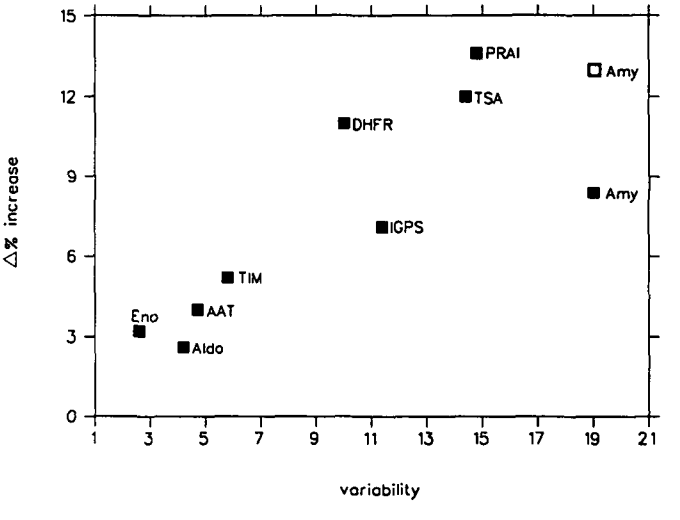


Fig. 3. Prediction accuracy improves with sequence variability. The difference of per-residue prediction accuracy between the average of single sequences and the averaged prediction of aligned sequences [see Δ% (av) in Table II] is plotted versus the overall average of the variability parameter as described in Materials and methods □, Improvement for Amy if optimal decision constants are used (see text).

Fig. 2. Secondary structure prediction is improved by weighting the averaged predictions with correlated averaged properties. (**A**) Section of aligned sequences of indoleglycerol phosphate synthase from 16 different organisms (α_2 to $\alpha_{8'}$) Organism acronyms. Eco, *Escherichia coli*; Sty, *Salmonella typhimurium*; Vpa, *Vibrio parahaemolyticus*; Bla, *Brevibacterium lactofermentum*; Ppu, *Pseudomonas putida*; Pae, *Pseudomonas aeruginosa*; Aca, *Acinetobacter calcoaceticus*; Rme, *Rhizobium meliloti*; Bsu, *Bacillus subtilis*, Lca, *Lactobacillus casei*; Sce, *Saccharomyces cerevisiae*, Ncr, *Neurospora crassa*; And, *Aspergillus nidulans*; Ang, *Aspergillus niger*; Pch, *Penicillium chrysogenum*; Pbl, *Phycomyces blakesleeanus*. Labelled rows: cons, conserved residues; single-letter amino acid code (upper case, conserved residues; lower case, partially conserved residues, allowing for two differences). X, Known secondary structure of the *E.coli* IGPS; wt, weighted averaged prediction; av, averaged prediction. Blank spaces in these lines correspond to gaps in the alignment. var, variability index (Wu and Kabat, 1970) for each position of the alignment, represented below as a truncated step profile. (**B**) Colour code: blue, helix and amphipathic moment profile; red, β-strand and hydropathy profile; green, coil and flexibility profile. **Panel A**, Profiles of three averaged properties; **panel B**, profiles of three averaged state propensities; **panel C**, the step profile, used for refined β-strand prediction as described in Materials and methods; **panel D**, profiles of three weighted averaged state propensities The β-strand profile is modulated further by profile of panel C as described in Materials and methods. The full-scale vertical axes in panels A, B and D correspond to ±400 centinats (Garnier *et al.*, 1978). The height of the steps in panel C corresponds to 100 centinats. Bar codes: X, wt and av as defined in (A).

$(\beta/\alpha)_8$-barrel enzymes (see Figure 2A). Therefore we used the pattern of sequence variability, to enhance specifically the prediction of β-strands, as follows. The weighted β-strand propensity profile—and only it—was incremented by a fixed value in regions where the average hydropathy was maximal and the average flexibility was minimal. This step function is shown in panel C of Figure 2(B). It is seen in panel D, that the previously unpredicted β-strands β_3 and β_8 were now predicted correctly. These observations were representative of the improvements, failures and remaining ambiguities found in the other seven proteins (not shown here).

The weighting procedure generally improved the prediction of all three states. The reason is that the property profiles correlate generally well with the known secondary structure elements. In particular, most peaks in the hydropathy profiles correlate better with known β-strands than peaks in the β-strand propensity profile. Moreover, all known β-strands correlate well with minima in the variability profile (see Figure 2A). They therefore all qualify for reinforcement of the β-strand prediction as described in Materials and methods. Figure 2(B) also shows that the underpredicted β-strand β_3 is associated with small amplitudes of the profiles of both averaged state propensity (panel B) and of averaged properties (panel A). Thus, on one hand, this segment (and similar segments in the remainder of the learning set) did not exceed the threshold required for applying the weighting procedure (see Materials and methods). On the other hand, the amplitudes of both the hydropathy and the flexibility profiles and the minimum in the variability profile associated with β_3 qualified for application of the reinforcement procedure.

The weighting procedure fails to improve the prediction in the region between residues 220 and 233; i.e. the loop around position 223 remains mispredicted as β-strand, and the region becomes even more ambiguous by insertion of a loop into the previously correctly predicted helix α_6.

Quantitative evaluation

To show how averaging and weighting of the secondary structure propensities increased the number of correctly predicted residue positions over those obtained from single sequences (si), we have (i) averaged the predictions of all the single sequences of a particular test protein, and (ii) compared that score to both the averaged (av) and to the weighted (wt) predictions.

To present an unbiased summary of the results we have followed Schulz' (1988) suggestion to use a scoring matrix (Table I). The numbers represent the sum of predicted residues from all seven test proteins divided by 7 to give a realistic impression of the error distribution observed with an 'average barrel enzyme', which has a total of 284 residues.

Table I shows that the number of predicted states increases from 'si' via 'av' to 'wt' for the β-strand and coil prediction, whereas the helix prediction remains practically unchanged. As shown in the off-diagonal cells of Table I, the increase of correctly predicted residues was necessarily paralleled by a decrease in some categories of mispredicted residues. Q_3, the total score of correctly predicted residues expressed in percent, was calculated from Table I as described in Materials and methods. It increases from single sequences (61%), via the averaging of the predictions of aligned sequences (68%), to the weighted predictions (71%). For comparison, the score of random assignment of secondary structure to proteins of the α/β class on a per-residue basis is 38% (Gibrat *et al.*, 1987; Schulz, 1988).

Table II summarizes the individual results obtained for all seven test proteins. The Q_3 values vary considerably amongst the members of the set, but the relative improvement (Δ%) obtained by averaging and weighting is not correlated to the Q_3 scores of the single sequences. $Q_{1,S}$ is the score of correctly predicted β-strand states expressed in percent. In proteins with highly diverse sequences as judged by the average variability criterion, the $Q_{1,S}$ value increases dramatically by averaging and weighting. These procedures lead to the recognition of previously unpredicted β-strands (see Figure 2B). This improvement is not adequately represented by the Q_3 value and requires an alternative scoring procedure.

Table I. Prediction score of three conformational states of an average $(\beta/\alpha)_8$-barrel enzyme on a per-residue basis

Method[a]	No. of predicted states[b]			Total of known states	
	Helix	Strand	Coil		
si	84	10	17		
av	87	11	13	111	Helix
wt	85	9	17		
si	16	20	11		
av	10	29	8	47	Strand
wt	4	36	7		
si	39	16	69		
av	24	22	78	126	Coil
wt	22	20	82		

[a]Method of prediction: si, single sequences, averaged, av, averaged; wt, weighted averaged.
[b]The results from the seven enzymes of the learning set were divided by 7 to yield results pertaining to an 'average' $(\beta/\alpha)_8$-barrel enzyme.

Score based on predicted secondary structure elements

Empirical prediction methods, which draw on a statistical data base, cannot be expected to predict the borders of structural elements exactly (Taylor and Thornton, 1984; Schulz, 1988). Moreover, the appropriate standard of comparison would have to be the sequence of average secondary structure elements determined from high-resolution crystallographic data on each member of the aligned set. It is unlikely that such data will become available in the forseeable future.

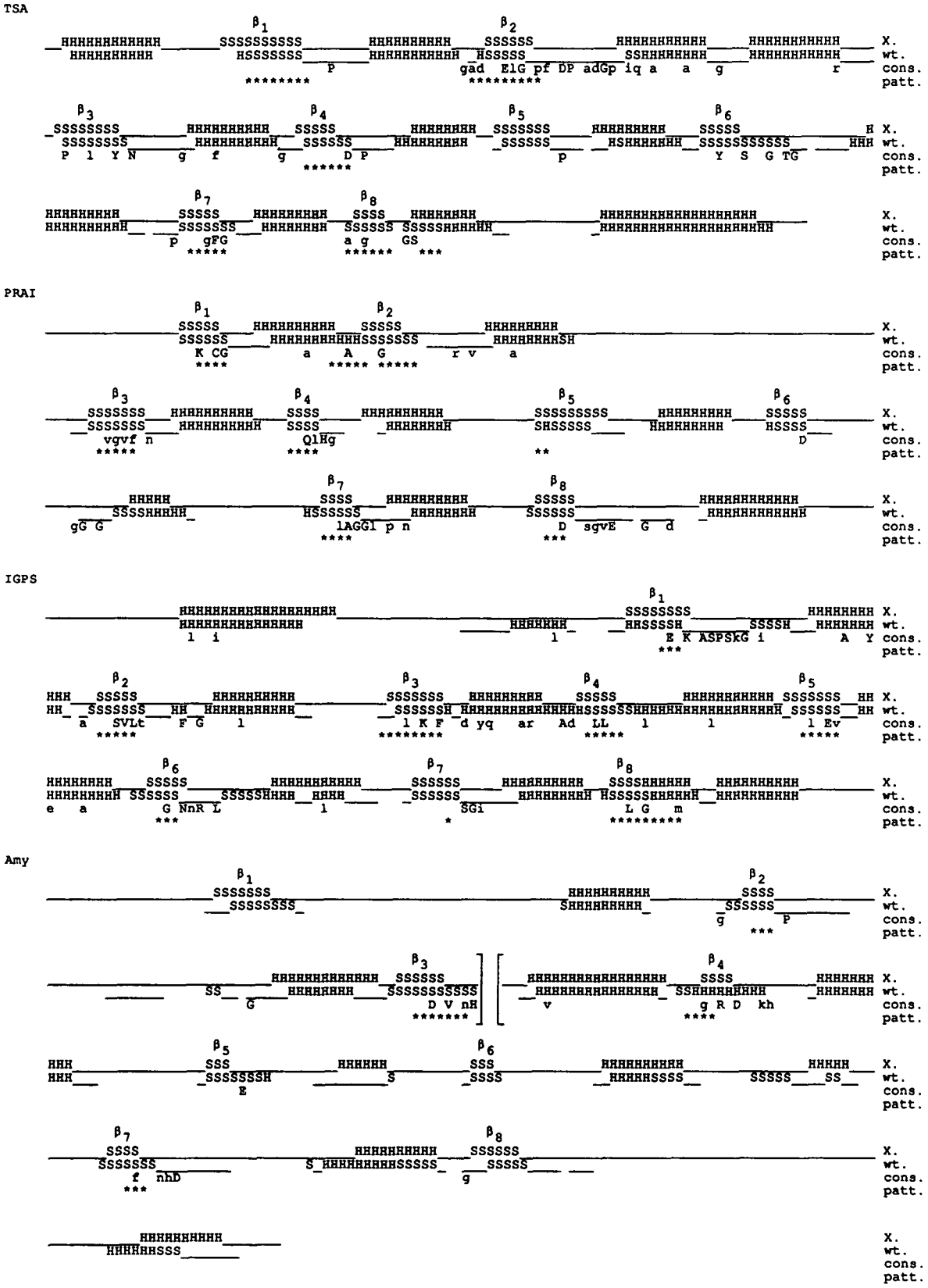
If secondary structure prediction is to be a step in recognizing the correct chain fold (Schulz, 1988), it is more important to predict correctly the number and sequence of secondary structure elements (here designated the 'secondary sequence'), rather than their exact borders. This presentation has the advantage that the 100% limit is clearly defined. Figure 4 presents the X-ray and the predicted secondary sequences of all seven $(\beta/\alpha)_8$-barrel enzymes. These pairs of sequences are abstracted readily for preparing the quantitative score (Taylor, 1984), as described in Materials and methods. The results obtained from the entire set of test proteins are presented in Table II.

The high accuracy of correctly predicted single states in the weighted prediction for TSA, PRAI and IGPS is more significant on a per-element basis than on a per-residue basis. An inspection of the per-element scores in the matrix cells of Table II reveals that the prediction of the secondary sequence of TSA, PRAI and IGPS is very close to 100% correct. This score is a significant improvement over the score of the simpler averaging procedure that led to the correct prediction of the chain fold of TSA (Crawford *et al.*, 1987). In the case of TIM three out of five false negative helices are 'loop helices' (see Figure 4). The part of the unorthodox $(\beta/\alpha)_8$-barrel chain fold of Eno was also recognized correctly.

In four cases the Q_3 values attain 75% correctly predicted single residues or higher. Taking these cases together, only one out of 38 known helices was missed in the prediction, and three predicted helices had to be classified as false positives. Only one out of 32 known β-strands was missed and five false positives were predicted.

Discussion

The availability of a relatively large number of homologous sequences of seven different $(\beta/\alpha)_8$-barrel enzymes prompted us to modify the existing methods of predicting the secondary structure for this predominant chain fold (Farber and Petsko, 1990) in the class of α/β proteins. Following the suggestion of



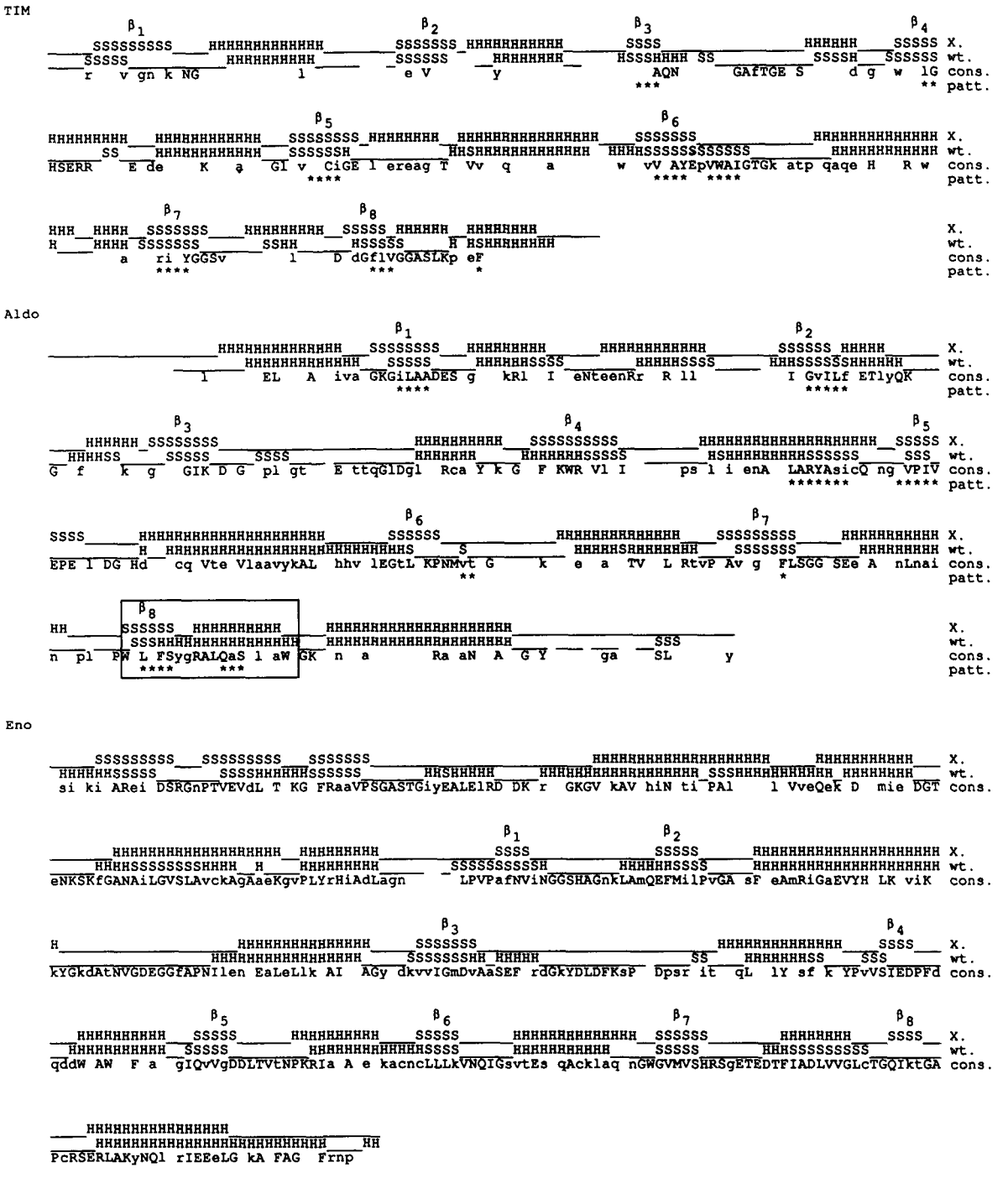


Fig. 4. Weighted state predictions of the seven $(\beta/\alpha)_8$-barrel enzymes (line 'wt'), aligned to the known secondary structure (line 'X'). The abbreviations of enzymes are given in footnotes to Table II. Totally conserved residues are indicated by their upper-case single-letter amino acid code. Positions where all except maximally two amino acids are conserved are indicated by lower-case amino acid codes (line 'cons'). Asterisks (line 'patt') indicate the occurrence of low sequence variability correlated to hydropathy and flexibility as described in Materials and methods. The pattern is omitted for Eno because of the generally low sequence variability in the set of aligned sequences. Parentheses between β_3 and α_3 of Amy indicate the position of domain B, which is no component of the $(\beta/\alpha)_8$-barrel fold. For the significance of the boxed region (β_8, α_8 of Aldo) see Materials and methods.

Garnier *et al.* (1978), it is shown that the averaged three-state prediction for all residues at aligned position (i.e. the per-residue Q_3 value, Table II) was significantly better than the average accuracy of that prediction for single sequences. The use of averaged property profiles, which correlate with the three conformational state profiles, as quantitative weights increased the accuracy to 67–82% on a per-residue basis. Use of the minima in the associated sequence variability profiles as quantitative reinforcement of the β-strand propensities improved the prediction even further, so that >95% of the known secondary elements were assigned correctly (see the per-element scores in Table II). This result constitutes a significant improvement over the previous joint prediction approach that led to the correct prediction of the chain fold of the α-subunit of TSA (Crawford *et al.*, 1987). Table II shows that only two out of 70 known elements were missed and eight were overpredicted.

Table II. Quantitative evaluation of secondary structure predictions—summary of data from seven $(\beta/\alpha)_8$ barrel enzymes

Enzymes[a] (no. of sequences)	Average variability[b]	Distribution of conserved residues[c]		Method[d]	Per-residue scores (% correctly predicted)					Per-element scores, weighted prediction[g]			
		C_S	C_H		$Q_{1,H}^e$	$Q_{1,S}$	$Q_{1,C}$	Q_3^f	Δ	H		S	
										tp	fp fn	tp	fp fn
TSA	14 4	6	0	si	76	50	62	66		11	0	8	1
(15)				av	87	74	70	78	12				
				wt	92	86	68	82	16		0		0
PRAI	14.8	5	0	si	75	51	57	62		8	0	8	1
(14)				av	81	82	67	76	14				
				wt	83	91	65	78	16		0		0
IGPS	11.4	7	1	si	80	32	46	58		10	1	8	2
(16)				av	94	43	46	65	7				
				wt	90	83	48	72	14		0		0
Amy	18.9	5	0	si	61	54	52	55		7	1	7	4
(21)				av	60	81	62	63	8				
				wt	60	78	68	67	12		2		1
TIM	5.8	7	2	si	67	43	53	58		7	2	8	1
(11)				av	66	62	60	63	5				
				wt	61	75	67	66	8		5		0
Aldo	4.2	8	4	si	78	33	60	63		12	1	7	4
(9)				av	75	47	64	66	3				
				wt	73	57	71	69	6		0		1
Eno	2.6	8	6	si	85	42	62	68		8	2	7	1
(7)				av	83	58	68	72	4				
				wt	81	68	69	75	7		1		1

[a]TSA, α subunit of tryptophan synthase; PRAI, phosphoribosyl-anthranilate isomerase, IGPS, indoleglycerol phosphate synthase, Amy, alpha amylase; TIM, triose phosphate isomerase; Aldo, aldolase; Eno, enolase. The number of individual amino acid sequences used is given below the name of the enzyme in parentheses.

[b]The variability parameter defined in Materials and methods averaged over the entire sequence.

[c]C_S, number of β-strand plus loop segments that carry conserved residues; maximum value 8. C_H, Number of α-helix plus loop segments that carry conserved residues; maximum value 8.

[d]Method: si, predictions of single sequences, averaged; av, averaged; wt, weighted averaged predictions of aligned sets of sequences.

[e]$Q_{1,H}$, $Q_{1,S}$, $Q_{1,C}$, percent of residues in state (H, S, C) predicted correctly as described in Materials and methods.

[f]Q_3, percent of total residues predicted correctly. Δ%, increase of Q_3.

[g]Per-element scores, based on weighted averaged predictions of aligned sets of sequences. The data of Figure 4 were evaluated by the structure abstract procedure described in Materials and methods. H, α-helix; S, β-strand. The score for each enzyme is presented as 2 × 2 matrix. See Materials and methods for definition of tp (true positive), fp (false positive) and fn (false negative).

Therefore, it is reasonable to conclude that a Q_3 value of 75% is sufficiently high to recognize the chain fold of a $(\beta/\alpha)_8$-barrel enzyme.

What are the possible reasons for the improvement achieved with the method described in this work? The symmetric chain fold of $(\beta/\alpha)_8$-barrel enzymes has a number of *a priori* favourable properties (see Figure 1): (i) it is a single domain; and (ii) medium-range interactions (amongst subsequent α-helices and β-strands) predominate, except for the packing of residues in the interior of the barrel (Lasters *et al.*, 1988; Lesk *et al.*, 1989).

Since the work of Wooton (1974) homologous sequence information has been used in various ways (Bajaj *et al.*, 1987; Fishleigh *et al.*, 1987; Webster *et al.*, 1987; Zvelebil *et al.*, 1987; Perkins *et al.*, 1988, 1989; Quian and Sejnowsky, 1988; Ghetti *et al.*, 1989) to improve the score of predicting secondary structure. Averaging *per se* and the smoothing procedures lead to a suppression of ambivalent regions of the state propensity profiles. Moreover, averaging takes into account the quantitative propensities of homologous secondary structure elements. Averaging is therefore preferable to consensus prediction procedures, which consider only the relative number of predicted states, but not their amplitudes. The chainfold of $(\beta/\alpha)_8$-barrel enzymes permits a structurally correlated assignment of certain physico-chemical properties (i.e. hydrophobicity and chain flexibility) with particular secondary structure elements, which focus on aspects of the protein structure that are different from directional information. The conserved sequence patterns which are characteristic for amphipathic α-helices, internal β-strands and neighbourhood correlations in oligopeptide loops at the surface seem to lead to the observed amplification of the corresponding propensities upon averaging the values at aligned residue positions. Thus, external α-helices are associated with amphipathic helical moments, internal β-strands with pronounced hydrophobicity and surface loops with high chain flexibility. However, some amino acids have ambivalent properties (e.g. the

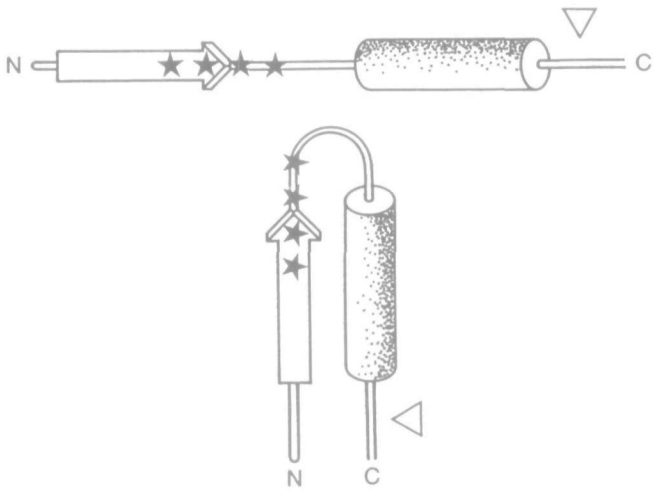


Fig. 5. Schematic representation of the repeated pattern of 8-fold $(\beta/\alpha)_8$-barrel enzymes. **Top**: linear array of β-strand (arrow) loop, α-helix (cylinder) and loop. N, N-terminus; C, C-terminus; ∇, sites that tolerate insertions and deletions; ★, conserved residues. **Bottom**: bending of the linear segment generates a $\beta\alpha$ supersecondary structure that is repeated 8-fold in $(\beta/\alpha)_8$ barrel enzymes (see Figure 1).

hydrophobic linker connecting the polar guanidino group of arginine to its α-carbon atom) and cannot be assigned uniquely to a particular secondary structure. Averaging of the property profiles of single protein sequences eliminates the ambiguity of individual amino acid residues. Similar to the effects observed with the state propensities, averaging also amplifies structurally correlated amplitudes and suppresses amplitudes which arise from ambivalent segments. The correlations between averaged propensity and property profiles are more pronounced than those of individual sequences. Thus the property profiles can be used to advantage for automatically weighting the secondary structure prediction.

Because all of the seven proteins are enzymes, there is incomplete separation between those regions that stabilize the fold and those regions that are responsible for the binding and turnover of substrates. This interdependence leads to preferred clustering of conserved residues in loops between β-strands and α-helices and of gaps in loops between α-helices and β-strands. Although this periodic pattern is not perfect it automatically subdivides the sequence into a number of segments (see Figure 2A). In addition, the lengths of α-helices and β-strands are rather narrowly distributed around the average values of α/β proteins determined by Taylor and Thornton (1984) (α, $n \simeq 12$; β, $n \simeq 5$).

The periodic occurrence of β-strands, α-helices, clustered conserved residues and gaps (see Figure 2) can be represented as a higher-level pattern that is characteristic of monomeric, single-domain $(\beta/\alpha)_8$-barrel enzymes. Figure 5 shows the following sequence of elements. First, a hydrophobic β-strand, frequently carrying clustered conserved residues at its C-terminus, is followed by a loop that also frequently carries clustered, conserved residues and tolerates insertions and deletions only rarely. Occasionally non-core helices are inserted into these loops (see $\alpha_{8'}$ in Figure 1). Second, an amphipathic α-helix is followed by a loop that frequently tolerates insertions and deletions. This sequence of elements is repeated eight times. Figure 5 also shows how bending of the first loop converts the repeating unit into a supersecondary structure. Circular packing of these units to form a central closed hyperboloid of eight parallel β-strands generates the barrel fold that carries the active site at the C-terminal end of the β-strands (Lasters *et al.*, 1988; Lesk *et al.*, 1989).

The higher-level supersecondary structure pattern of Figure 5 can be used to detect and correct most of the mis- or unpredicted secondary structure elements shown in Figures 2 and 4. Additional cause for reconsideration is given in the case of multidomain or oligomeric $(\beta/\alpha)_8$-barrel enzymes (e.g. TIM, Aldo and Eno) where the various domains or subunits frequently interact via α-helices, superimposing conservation restraints that are extrinsic to the stability requirements of limit $(\beta/\alpha)_8$-barrel enzymes *per se*. Therefore the pattern can be used to discriminate tentatively between different possible chain folds in cases where the predicted pattern is not as clear-cut as for PRAI, IGPS and the α-subunit of TSA (see Figure 4).

It is interesting in this regard that the secondary structure of Eno, a heterodox 8-fold barrel enzyme (Lebioda *et al.*, 1989), is also predicted with high accuracy on a per-element basis. In particular, the unusual N-terminal sequence $\beta_1\beta_2\alpha_1\alpha_2$ is correctly predicted, probably because the environments of the antiparallel β-strand β_2 and the associated helix α_2 are identical to those of the other secondary structure elements.

Can the general approach that has been developed for known $(\beta/\alpha)_8$-barrel enzymes be applied to other proteins? A problem might arise in the alignment of a protein with unknown structure. The correct location of gaps is important because, in the final version of the alignment, even a single gap leads to a loop prediction. With proteins of unknown structure it might be difficult to assign gaps in an unequivocal manner especially if the sequence variability in that region is high. It may become necessary to reconsider the alignment where gaps and predicted α-helices or β-strands overlap. For example, if a gap interrupts a strong peak in the α-helix profile, it must be shifted to one of the flanks of the peak. In practice only the borders of an aligned gap are somewhat arbitrary, but not its location between two predicted secondary structure elements. This statement is supported by the correct location of all five clustered gap positions in the aligned sequences of TSA before the structure was known (Crawford *et al.*, 1987). It should be possible to determine from inspection of the two aligned sets of averaged propensity and property profiles whether an unknown protein belongs to the α/β class (Levitt and Chothia, 1976) or not. Averaging improves the predicted conformational propensity also of some other chain folds of the α/β class, e.g. DHFR and AAT (see Figure 3). The averaged property profiles can also be used to enhance qualitatively the tentative prediction of secondary structure elements of α/β proteins. Due to their different chain folds it will be difficult to identify β-strands that are at the edge of extended β-sheets and are therefore not hydrophobic, or internal α-helices, which are not amphipathic. It is therefore not advisable to use the automatic weighting or the reinforcement procedure of β-strands that is based on minima of sequence variability.

The strongly predicted propensity peaks that correlate well with the cognate property peaks can be recorded as relatively reliable regions of secondary structure, using the rules of Taylor's (1984) structure abstract procedure. Sometimes a strong peak in the property profile may enhance a weak peak in the propensity profile, and *vice versa*. The tentative sequence of secondary structure can then suggest the use of *ad hoc* templates for enhancing ambiguously predicted regions (Webster *et al.*, 1987). Furthermore, relevant information from genetics, the effects of chemical modification and limited proteolysis (Crawford *et al.*, 1987; Hurle *et al.*, 1988) can, in principle, help to support the prediction.

Acknowledgements

We thank M.Wilmanns and J.N.Jansonius for the refined secondary structure assignments of PRAI:IGPS and AAT, and C.Chothia for encouragement. This work was supported by the Swiss National Science Foundation, grant no. 3.255-0.85.

References

Argos,P. (1990) *Methods Enzymol.*, **182**, 751–775.
Argos,P. and McCaldon,P. (1988) In Setlow,J. and Hollander,A. (eds), *Genetic Engineering*. Plenum Press, New York, Vol. 10, pp. 21–66.
Argos,P., Schwarz,J. and Schwarz,J (1976) *Biochim. Biophys. Acta*, **439**, 261–273.
Bajaj,M., Waterfield,M.D., Schlessinger,J., Taylor,W.R. and Blundell,T. (1987) *Biochim. Biophys. Acta*, **916**, 220–226.
Biou,V., Gibrat,J.F., Levin,J.M., Robson,B. and Garnier,J. (1988) *Protein Engng*, **2**, 185–191.
Chothia,C. and Lesk,A.M. (1987) *EMBO J.*, **5**, 823–826.
Chou,P.Y. and Fasman,G.D. (1978) *Adv. Enzymol.*, **47**, 45–148.
Cohen,F.E., Abarbanel,R.M., Kuntz,I.D. and Fletterick,R.J. (1983) *Biochemistry*, **22**, 4894–4904.
Cornette,J.L., Cease,K.B., Margalit,H., Spouge,J.L., Berzofsky,J.A. and DeLisi,C. (1987) *J. Mol. Biol.*, **195**, 659–685.
Crawford,I.P., Niermann,T. and Kirschner,K. (1987) *Proteins*, **2**, 118–129.
Eichele,G., Ford,G.C., Glor,M. and Jansonius,J.N. (1979) *J. Mol. Biol.*, **133**, 161–180.
Eisenberg,D., Weiss,R.M. and Terwilliger,T.C. (1984) *Proc. Natl Acad. Sci. USA*, **81**, 140–144.
Farber,G.K. and Petsko,G.A. (1990) *Trends Biochem. Sci.*, **15**, 228–234.
Fishleigh,R.V., Robson,B., Garnier,J. and Finn,P.W. (1987) *FEBS Lett.*, **214**, 219–225.
Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **120**, 97–120.
Genetics Computer Group (GCG), University of Wisconsin, Sequence Analysis Software Package, Version 6.1 (1989).
Ghetti,A., Padovani,C., DiCesare,G. and Morandi,C. (1989) *FEBS Lett.*, **257**, 373–376.
Gibrat,J.-F., Garnier,J. and Robson,B. (1987) *J. Mol. Biol.*, **198**, 425–443.
Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
Hurle,M.R., Matthews,C.R., Cohen,F.E., Kuntz,I.D., Toumadje,A. and Johnson,W.C. (1987) *Proteins*, **2**, 210–224.
Hyde,C.C., Ahmed,S.A., Padlan,E.A., Miles,E.W. and Davies,D.R. (1988) *J Biol. Chem*, **263**, 17857–17871.
Jaenicke,R. (1987) *Prog. Biophys. Mol. Biol.*, **49**, 117–247.
Kabsch,W. and Sander,C. (1983) *FEBS Lett.*, **155**, 179–182.
Karplus,P.A. and Schulz,G.E. (1985) *Naturwissenschaften*, **72**, 212–213.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
Lasters,I., Wodak,S.J., Alard,P. and van Cutsem,E. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 3338–3342.
Lebioda,L., Stec,B. and Brewer,J.M. (1989) *J. Biol. Chem.*, **264**, 3685–3693.
Lesk,A.M., Brändén,C.-I. and Chothia,C. (1989) *Proteins*, **5**, 139–148.
Levitt,M. and Chothia,C. (1976) *Nature*, **261**, 552–557.
Lim,V.I. (1974) *J. Mol. Biol.*, **88**, 873–894.
Matthews,B.W. (1975) *Biochim. Biophys. Acta*, **405**, 442–451.
Matthews,D.A., Alden,R.A., Bolin,J.T., Freer,S.T., Hamlin,R., Xuong,N., Kraut,J., Poe,M. and Hoogsteen,K. (1977) *Science*, **197**, 452–455.
Mewes,H.W. (1990) MIPS; Max-Planck Institute for Protein Sequences.
Nagano,K. (1973) *J. Mol. Biol.*, **75**, 401–420.
Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
Perkins,S.J., Haris,P.I., Sim,R.B. and Chapman,D. (1988) *Biochemistry*, **27**, 4004–4012.
Perkins,S.J., Nealis,A.S., Dudhia,J. and Hardingham,T.E. (1989) *J. Mol. Biol.*, **206**, 737–753.
Priestle,J.P. (1988) *J. Appl. Crystallogr.*, **21**, 235–248.
Priestle,J.P., Grütter,M.G., White,J.L., Vincent,M.G., Kania,M., Wilson,E., Jardetzky,T.S., Kirschner,K. and Jansonius,J.N. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 5690–5694.
Quian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.
Schulz,G.E. (1988) *Annu. Rev. Biophys. Biophys. Chem.*, **17**, 1–21.
Strasser,A.W.M., Selk,R., Dohmen,R.J., Niermann,T., Bielefeld,M., Seeboth,P., Tu,G. and Hollenberg,C.P. (1989) *Eur. J. Biochem.*, **184**, 699–706.
Taylor,W.R. (1984) *J. Mol. Biol.*, **173**, 512–514.
Taylor,W.R. (1987) In Bishop,M.J. and Rawlings,C.J. (eds), *Nucleic Acid and Protein Sequence Analysis: a Practical Approach*. IRL Press, Oxford, pp. 285–322.
Taylor,W.R. (1988) *Protein Engng*, **2**, 77–86.
Taylor,W.R. and Thornton,J.M. (1984) *J. Mol. Biol.*, **173**, 487–512.
Webster,T.A., Lathrup,R.H. and Smith,T.F. (1987) *Biochemistry*, **27**, 6950–6957
Wooton,J.C. (1974) *Nature*, **252**, 542–546.
Wu,T.T. and Kabat,E.A. (1970) *J. Exp. Med.*, **132**, 211.
Zvelebil,M.J., Barton,G.J., Taylor,W.R. and Sternberg,M.J.E. (1987) *J. Mol. Biol.*, **195**, 957–961.

Received on June 19, 1990; revised and accepted on September 21, 1990